



EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

**A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND
APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES**

- Clusters from Iterative, Intercolumnar Correlational Analysis.* LOUIS L. MCQUITTY AND JAMES A. CLARK 211
- Tetrachoric Correlation by the Camp Approximation.*
EDWARD E. CURETON 239
- A Measure of the Average Intercorrelation.* HENRY F.
KAISER 245
- An Important Similarity between Biserial r and the
Brogden-Cureton-Glass Biserial r for Ranks.* JULIAN
C. STANLEY 249
- Estimating Individual Rater Reliabilities from Anal-
ysis of Treatment Effects.* JOHN E. OVERALL 255
- An Application of Balanced Incomplete Block Designs
to the Estimation of Test Norms.* THOMAS R. KNAPP 265
- The Interrelationships among Item Characteristics in
an Adjective Check List: The Convergence of Dif-
ferent Indices of Item Ambiguity.* LEWIS R. GOLD-
BERG 273

(Continued on inside front cover)

(Continued from front cover)

<i>Item Quality and Appropriateness of Response Processes.</i> CASTELLANO B. TURNER AND DONALD W. FISKE	297
<i>Prediction of Anchor Effects on Personality Items from Rating Dispersions.</i> JAMES A. WALSH	317
<i>Medical Interns: A Study of Types.</i> MARGARET A. HOWELL ...	327
<i>Social Intelligence and IQ.</i> RALPH HOEPLNER AND MAUREEN O'SULLIVAN	339
<i>An Exploratory Study of Programmed Tests.</i> T. ANNE CLEARY, ROBERT L. LINN, AND DONALD A. ROCK	345
<i>Intellectual Interest Patterns of Gifted College Students.</i> JOSEPHINE S. GOTTSACKER	361
<i>Stability of Engineering Interests over a Period of Four Years.</i> W. SCOTT GEHMAN AND ILA H. GEHMAN	367
VALIDITY STUDIES SECTION	377
BOOK REVIEWS	619

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Contributors receive one hundred reprints of their articles without charge. Manuscripts should be sent in duplicate to G. Frederic Kuder, Box 6907, College Station, Durham, North Carolina 27708. Authors are requested to put tables and footnotes on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1957 Revision)*. Journal titles should not be abbreviated.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at 2901 Byrdhill Road, Richmond, Virginia 23205. Second class postage paid at Richmond, Virginia.

Subscription rate, \$14.00 a year, domestic and foreign. Single copies, \$3.50. Back volumes: Volumes XX to the present \$14.00 each; Volumes V through XIX, \$10.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27708.

THE INTERPRETATION OF WRONG ANSWERS FROM A MULTIPLE CHOICE TEST

J. C. POWELL

University of Alberta

WITH an ever-increasing emphasis in our schools and clinics on the individualization of treatment, there are several aspects of our diagnostic procedures which require re-examination. The multiple choice test is a useful general screening device. However, for more detailed information we still resort to prolonged personal contact, often at considerable expense of time and effort. Is it possible that the precision of our screening devices can be improved to facilitate diagnosis and hence prognosis and treatment?

This paper will concern itself with one limited aspect of this problem; namely, the functional role of the "wrong answers" on multiple choice tests. Much time is spent by the examiner in the preparation of foils for multiple choice tests. A proportionally large time is spent by the examinee in making his selection decisions among the alternatives. In spite of the time thus spent, the foils are generally treated as a mask to the right answer and are lumped together in a general "wrong category." The rating of the examinee is usually entirely dependent on his total number of correct items on any given test or subtest. On the other hand, if a multiple choice test has been well prepared, particular "wrong" answers may have nearly equivalent discriminating power as do the "right" answers.

If foils do not have discriminating power other than to act as a mask for the "right" answer, then it is reasonable to assume that the selection of these foils by the examinee should be roughly random in distribution. This randomness assumption suggests several hypotheses. First, if examinees are asked to give reasons for

their selection, there should be little more than chance equivalence among their stated reasons on any particular foil selection for all those individuals who select this particular foil.

Second, although some foils may hang together factorially between items, there should be no logical consistency between stated reasons within any particular statistical factor.

Third, if examinees were to be subdivided into two groups, there should be little if any consistency between the reasons given by the members of the group of examinees on whose responses the statistical determinations were made and the stated reasons for the response selections of the second group of examinees on these same foils.

Fourth, if two or more individuals choose the same pair of foils, that is, both individuals make identical errors on at least two different items, we will call this equivalence *overlap*. To maintain the random assumption, there should be no relationship between the proportion of overlap of each individual's "errors" with the errors of all other individuals and their rank on the total-correct score which they receive.

Fifth, if we were to subdivide the examinees into five groups on the basis of percentile rank with respect to total-correct score, the random assumption suggests that for all such groups, the selection of all foils on any item generally should not be significantly different from random.

Sixth, although some of the distributions in case five will be significant, there should be no consistent similarity between the placement of these significant deviations from random with the other wrong answer variables in the same factor.

Although these six relationships do not exhaust the possibilities which the assumption of the random selection of foils would predict, these are sufficient that if a substantial proportion of these conclusions are refuted, the random assumption is also refuted.

Procedure. In order to test these six hypotheses a commercially prepared multiple-choice test was used. This test, The Proverbs Test, by D. R. Gorman (1957) was used for two reasons. First, we did not wish to use a test which might reflect the particular bias of the author. Second, the test was developed from the scoring of a projective version of the same test. This means that the foils are based on the actual responses of disturbed persons and should, for

this reason, reflect random selection if any foils set is going to do this.

The test was administered to two classes of upperclassmen in a course in educational measurement. The members of one of these two classes ($N=18$) will be called Group A. The members of the other class will be called Group B ($N=23$).

The following analyses were conducted:

1. A binary matrix of the correct responses to every item by every subject in Group A was truncated to remove items with zero or near zero variance. The remaining matrix was factor analysed.
2. A binary matrix of the particular foil selections among each of the three possible wrong responses to every item by every subject was also truncated to remove variables of zero or near zero variance. The remaining matrix was factor analysed. The criterion for truncation in both these instances was based on the response ratio (D):

$$D = \sum_{i=1}^N X_{i,j}/N$$

where $X_{i,j}$ is the binary value of the selection of the j th alternative by the i th individual.

Such that $(.10 \leq D \leq .90)$.

3. Those foils which were related factorially were examined on the basis of the stated reasons for their selection to determine the possibility of consistency of these reasons within particular foils, and among factorially related foils. This examination was used to study the problems arising from the first two hypotheses mentioned above with reference to the random assumption for foil selection.
4. The most frequently stated reason for the selection of a retained wrong answer in Group A became the variable which was used to predict that a semantically equivalent reason would be given for the same foil by anyone in Group B who selected it. A binary vector was developed equating the "hits" and "misses" with the total possible hits. This proportion was treated as an estimate of the variance accounted for in cross-validation. This procedure was designed to examine the third hypothesis.

5. The average *overlap* (as defined above) of each examinee's foil selection was calculated and the examinees were then ranked on the basis of descending order of these values. A rank order correlation coefficient was then calculated between rank order of overlap and the corresponding rank order by total score correct for each examinee. This procedure refers to the fourth hypothesis.
6. Groups A and B were combined and subdivided on the basis of rank order by total score correct with roughly one fifth of the combined group of examinees in each of these new subdivisions. Thus "UPPER 5" (see Table 3) refers to those individuals who received total correct scores in the top 20 per cent of the entire group. A *chi square* test for independence was done across all the foils of every item which had one of its foils as one of the retained foil variables. This procedure was used to examine the last two of our random assumption hypotheses.

Results. The results of step one of the analysis left 32 of the 40 items for factoring. A principal components analysis of these gave a very clear simple structure without rotation. There were three factors with *eigenvalues* greater than 1.00 with the first factor containing 29 of the items. When these 29 variables were used they represented a unifactor which accounted for 81.4 per cent of the variance on the test as shown in Table 1. The nature of this test suggests that the reduced test is a "Translation" test. This would put it into the comprehension level of Bloom's (1956) *Taxonomy*.

The foil selection pattern was subjected to a similar procedure. In this case there was no clear pattern in the original factor matrix. For this reason several *varimax* rotations were performed on different numbers of factors. The criterion sought was the highest number of variables with high loadings on any one factor and low loadings on all others. A four factor rotation gave the best results to this criterion. The number of variables then considered was reduced to those which showed simple structure. Variables among wrong answers were also rejected if their right answer had been rejected in step one. This procedure left 15 foils distributed across the 29 items which had a total communality of 9.115 which when distributed across $3 \times 29 = 87$ possible

foils accounts for 10.5 per cent of the variance as is shown in Table 1. Thus the combined results of both the right and the wrong answers for the shortened 29 item test accounts for a total of 91.9 per cent of the variance on this test.

TABLE 1
Factor Patterns for Right and Wrong Answers

Right Answers—First Principal Axis Hotelling Solution Test Items Retained		Wrong Answers Varimax Rotation Alternatives Retained				
Item ^a	Factor Loading	Item ^a	I	II	III	IV
1c	.929	3a	.590			
2d	.898	14c	.732			
3d	.938	22a	.900			
6c	.941	27a	.518			
8c	.954	29b	.765			
		13d		.916		
		15d		.647		
9b	.839	19c		.916		
10c	.835	23d		.668		
11d	.924	9c			.814	
13c	.838	13a			.682	
14d	.944	18a			.689	
		2b				.713
		20a				.744
15c	.905	21a				.619
17a	.966					
18b	.948					
19d	.918					
20c	.904					
		Alternatives Dropped				
		10a		.442		
		16a ^b			.874	
21c	.898	16b	.661			.416
22c	.914	23a			.567	
23c	.849	24d		.442		
24c	.935	25a ^b	—	—	—	—
26b	.935	25c				.801
		31d	.594	.526		
27b	.942					
29c	.886					
31b	.897					
32a	.911					
34c	.935					
35c	.841					
36b	.954					
39b	.855					
40c	.848					
Total Communality 23.708		Total Communality of Retained Alternates 9.115				

^a This number and letter combination refers to corresponding item and alternative on the *Proverbs Test*.

^b Items 16 and 25 were eliminated from test on right answer reduction.

A careful examination of the reasons given for the selection of the foils for each of the 15 foil variables retained showed that about 60 per cent of the examinees gave semantically equivalent explanations for their selection.

When these reasons were examined for the logic behind them across all the members of a foil factor, a fairly clear psychological interpretation of each factor occurred. These factors displayed the following possible interpretation:

- Factor I: In each of these cases the subject seems to be simplifying part of the information contained in the proverb.
- Factor II: In this case each of the subjects seems to be adding an additional bit of irrelevant information in arriving at the answer.
- Factor III: In this case there seems to be a substitution of meaning for one element in the answer.
- Factor IV: In this case the selected foil seems to be another proverb bearing little or no relationship to the proverb stated in the stem.

These factor interpretations seem to be consistent with a logical analysis of the items themselves. Table 2 illustrates this point by selecting one item from each sub-test, giving the appropriate foil for the sub-test, and listing examples of the reasons for selection, from both groups of examinees.

When Group A's usual reasons were used to predict the reasons for foil selection given by Group B, it was found that Group B's reasons were semantically equivalent to the most frequent reasons in Group A 64 per cent of the time. This finding would imply a cross-validation correlation coefficient of $r = \sqrt{.64} = .80$. (Notice the equivalence indicated in Table 2.)

The rank order correlation between the proportion of overlap and the rank order of total correct score was .47 which is significant at the .05 level. This finding implies that the errors of the higher performing students are more alike than those of the lower performing students. This situation is, of course, also true for their correct responses.

Finally, a chi square test of independence on the ranked groups of the combined classes, showed that a proportion of the foil

TABLE 2

Sample Analysis of Items to Illustrate Classification of Foils

Factor I	Reduction of information to affect simplification of statement.
Example	Item 22. A ROLLING STONE GATHERS NO MOSS. Foil a. BE CONSISTENT.
Analysis	This stem can be described as a relationship between an object an action, and a negative condition. The foil drops the negative aspect of the condition, changing the meaning. BE INCONSISTENT would be better, but in this case, the relative value judgement of "moss gathering" in the original is reversed to the usual connotation given this proverb.
Sample Responses	
First Group	1. Not settled—rolling stone. 2. Related rolling stone with consistent.
Second Group	1. Rolling stone—don't settle down. 2. Consistent and no moss.
Comments:	In each case the reason listed seems to be more closely related to the dropping of the negative than misunderstanding the proverb. In other words, this error is in the translation process rather than in the comprehension of the stem.
Factor II	Addition of irrelevant information.
Example	Item 13. THE WORST SPOKE IN A CART BREAKS FIRST. Foil d. IT TAKES A GOOD MAN TO KEEP ON TRYING.
Analysis	Here we have an object-event-condition statement again. The object is treated as a unit. In the foil this unitary characteristic is changed and the weak point isolated as separate from the whole, a value judgement is then added.
Sample Responses	
First Group	1. A weak person will stop trying.
Second Group	2. No interpretable reasons in this group.
Comment:	Again the translation process seems to be the source of the error.
Factor III	Substitution of elements.
Example	Item 9. THE MORE COST, THE MORE HONOR. Foil c. THE HIGHER THE PRICE, THE BETTER A THING IS.
Analysis	In this case "honor" is translated as "better" rather than as "appreciate" in the correct response. This subtle shift in meaning leads to a statement about the quality of an object itself rather than the value attributed to it. Hence, we classify this error as being the product of inappropriate substitution in the translation process.
Sample Response	
First Group	1. Cost implies price.
Second Group	2. Parallel parts in both elements.
Comment:	By restricting the meaning of the word "cost" when translating a substitution for "honor" becomes necessary. There is a degree of misunderstanding of the stem in this example.

TABLE 2—Continued

Factor IV	Replacement of proverb by one largely unrelated (Irrelevancy Fallacy).
<i>Example</i>	Item 20. THE USED KEY IS ALWAYS BRIGHT. Foil a. SOMETHING OLD IS BETTER THAN SOMETHING NEW.
<i>Analysis</i>	The term "used", used here in the sense "much used" is translated by substitution into "old", and the term "always bright" into "better than". Hence the end product has no similarity in meaning to the stem.
<i>Sample Response</i>	
<i>First Group</i>	1. Implies sentiment for old things as does stem.
<i>Second Group</i>	2. The word "used" key to answer.
<i>Comment:</i>	In this factor we can argue with some safety that the errors arose from a failure to understand the stem. It is interesting to note that a clear lack of understanding accounts for <i>less than two per cent</i> of the variance on this test.

selection which is different from chance is considerably greater than would be expected if foil selection were randomly distributed. The .10 level of significance was used because with a combined N of 41 which is then subdivided into five groups of about eight each, such small numbers are involved that a more stringent test would be unreasonable. Table 3 recapitulates the results. It should be noted that the upper 5 approached .10 level for every case in Factor IV. The small number of cases in this "Irrelevancy" factor have an adverse effect on the possibility of obtaining stable significant results for this factor.

Discussion. All six of the hypotheses derived from the assumption of random selection of foils must be rejected by the finding of this study.

First, not only are the reasons people give for selecting a foil consistent for that particular foil, but they are also consistent across members of the same statistically determined factor and also in a cross-validation with another group.

Second, the responses of examinees tend to become progressively more alike in both right and wrong responses as their total correct score increases.

Third, not only are particular foils inclined to identify particular groups within specific percentile ranges, but the identification of these groups is consistent across all members of the same factor and relatively distinctive from other factors. It is probable

TABLE 3

Significance of Foil Selection for Wrong Answer Factors^a

FACTOR I					
Level of Total Score	3a	14c	FOILS 22a	27a	29b
Upper 5	— ^b	—	—	—	—
4	—	—	—	—	—
3	.05	.05	.10	.05	—
2	.02	.02	.01	.02	.01
1	—	—	—	—	—
FACTOR II					
	13d	15d	19c	23c	
Upper 5	—	.10	—	—	
4	—	—	—	—	
3	—	—	—	—	
2	.05	—	.10	—	
1	.05	.10	.10	.05	
FACTOR II					
	9c	13a	18a		
Upper 5	—	—	.10		
4	—	—	—		
3	—	.10	—		
2	—	—	—		
1	.01	.10	.10		
FACTOR IV					
	2b	20a	21a		
Upper 5	—	—	.10		
4	—	—	—		
3	.10	.02	.10		
2	.10	—	—		
1	.10	—	—		

^a Chi square test for independence.^b Only p of $\chi^2 \leq .10$ are shown.

that this last finding identifies the basis upon which the factor analysis separates the factors. If this is the case, then this procedure may make it possible to differentiate between individuals with the same total-correct score on the basis of differential foil selection.

On this basis, we can state that the random assumption for the selection of foils for this particular multiple-choice is not tenable. Because of the nature of this test, this statement could have a broad generality.

If foils on multiple choice tests contain information about the examinees, what is the value of this information in a clinical or educational setting for purposes of diagnosis, prognosis, treatment,

or determining the effectiveness of treatment? Any statement in this area would be unduly speculative at the moment. Attempts to answer these questions might open fruitful avenues of enquiry.

Summary. A commercially prepared multiple choice test, The Proverbs Test, was analyzed to determine the possible information content of the wrong answers. It was demonstrated that the assumption that wrong answers are randomly distributed does not hold in this case. Wrong answers clearly have psychological and statistical meaning. The implications of these findings could have great importance for test interpretation. Further research is necessary to define these implications.

REFERENCES

- Bloom, B. S. (Ed.). *A Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: David MacKay, 1956.
Gorham, D. R. *The Proverbs Test*. Missoula, Montana: Psychological Test Specialists, 1957.



